NLP Modeling and Analysis of Popular Vloggers

Zhaofang(Zach) WANG
IPHS 300 Kenyon College

Introduction

Vlog, or video blog, is the chicest form of personal production on YouTube. First appearing at the beginning of the 21st century, vlogs saw an enormous rise in popularity around 2005. Differing from other forms of production, vlogs allow their producers more freedom in expressing themselves without the limitation of content, word count, and forms of expression. The absence of censorship and early-stage marketing allows an almost free-market effect in the sense that the most popular vloggers and videos are purely the choice of people. Therefore, analyzing viral vlogs and vloggers can offer a more accurate representation of where the general population's interests lie.

In the project, I focused on 4 vloggers of different genres to analyze the reasons behind their successes using NLP techniques.

Data

For the purpose of the study, four popular vloggers with different target audiences and styles were selected. The goal of the selection was to cover as many age and gender groups as possible, and the four vloggers are the following:

1. Logan Paul
2. Joana Ceddia
3. Colleen Vlogs
4. Emma Chamberlain

To ensure a non-biased dataset, the actual videos of each vlogger were chosen using a random number generator from a population of each vlogger's most popular videos with subtitles. Additionally, to make sure the lengths of videos being analyzed are roughly the same among the vloggers, I set a goal of 100 mins with a 10 mins margin, that is to say the total duration of videos of each vlogger being included in my corpus all fell in the interval of 90 mins -110 mins.

After 2 rounds of data cleaning, a brief summary of the data is as follows:
*The lengths of the videos are rounded to the nearest integer.

	Word Count	Video Length
Logan Paul	7908	103
Joana Ceddia	17950	99
Colleen Vlogs	17433	108
Emma Chamberlain	17232	102

Methodology and EDA

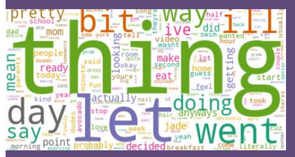
The main tools being used to tackle the problem were NLTK, Textblob, WordCloud, and Pandas, all of which are free python libraries for data and text analysis. Among the aforementioned tools, NLTK and Textblob were used for the purpose of text cleaning, data visualization, and sentiment analysis; Pandas was used for text cleaning and tokenization; WordCloud was used for data visualization.

Because the main interest was to explore the potential characteristics that could contribute towards the four vloggers' successes, and to see how the differences in target audiences influence the style of presentation, a visualization of most frequent words of each vloggers is presented below:

Colleen Vlogs

Emma Chamberlain

Joana Ceddia



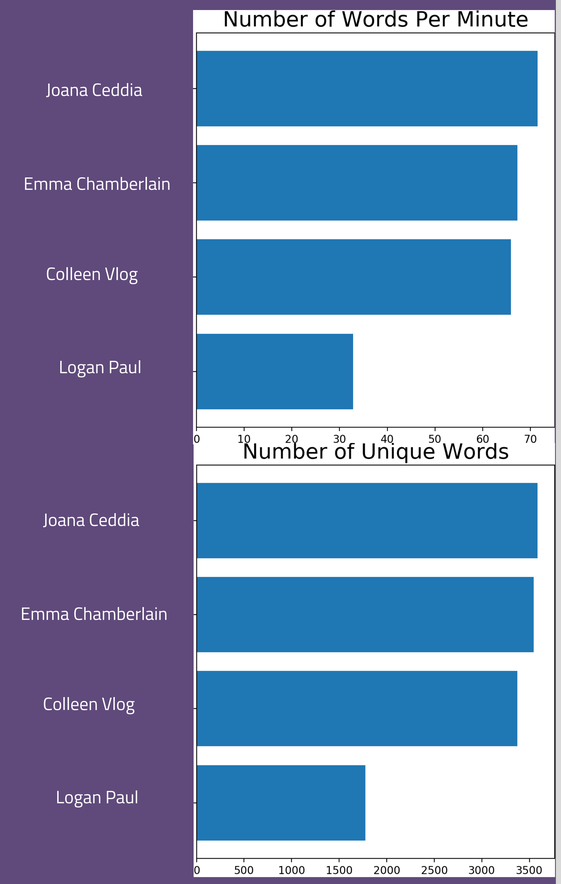
Logan Paul

From the word clouds, it is evident that differences in main topics and styles of presenting are present: "baby" appears to be one of the major topics of Colleen Vlogs; Emma Chamberlain swears quite often. While Joana Ceddia and Logan Paul seem to be lacking concrete topics, the style of presenting can still be inferred from the most frequent words.

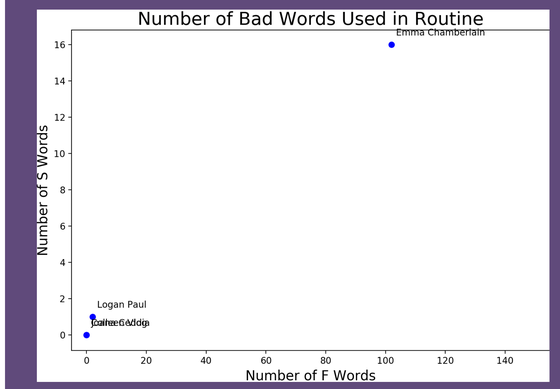
Further Exploration of Data

A very important aspect of vlogs is the style of presenting. A common trade-off for vloggers is often content vs. speech. Different groups of audience often have different preferences regarding the balance between content and speech. Therefore, intuitively it is essential for a vlogger to use a method of presenting that is the most preferred by their target viewers. A histogram serves well for the visualization of the problem.

Style of Presenting

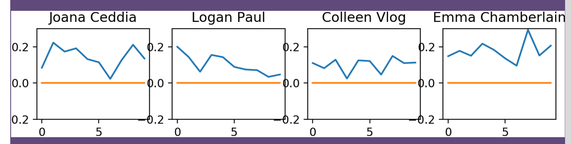


Visibly, vocabulary of Logan Paul's videos is much smaller comparing with the rest of the vloggers. The graph below shows the difference in the use of profanity.



Polarity Analysis

Another important feature of vlogs is polarity, meaning the positivity of the vloggers. A value below zero indicates an overall negative impression while a value above zero indicates optimism. To analyze the change in polarity over time, the corpus were divided into 10 subsections. The change in polarity of the four vloggers are plotted as follows:



While it is not rare to observe an overall positive sentiment, it is quite shocking to notice that the polarity of all four vloggers has never dropped below 0. That is to say, not only their videos are positive overall, there is not a single subsection where pessimism becomes dominant.

Conclusion

The results of data visualization mainly concur with the initial prediction. Logan Paul, with its primary viewers being of age 10-18, he uses more visual content and considerably fewer words. Colleen Vlog talks more about her new-born baby and also uses no profanity in her videos, considering her main audience being young female adults. Emma Chamberlain uses a shocking amount of profanity as her "cool" personality has always been a appeal to her target audience: rebellious teenagers. However, it remains unclear from the analysis why Joana Ceddia has risen to fame in a short span of time.

As the project merely focuses on the text aspect of vlogs while visual and audio content are also crucial parts of videos. Further studies on the other aspects can definitely render the project more comprehensive and accurate.

Reference

https://socialblade.com/

Zhao, Alice. (2019, August 14). adashofdata/nlp-in-python-tutorial. Retrieved from https://github.com/adashofdata/nlp-in-python-tutorial.